图书情报工作
第 62 卷 第 4 期 2018 年 2 月

中文科研论文未被引探索Ⅱ:基于关键词的内容因素影响研究
——以图书馆情报与文献学为例

韩毅 伍玉 申东阳 况书梅 袁庆
西南大学计算机与信息科学学院 重庆 400715

摘要: [目的/意义]从内容差异来探索论文未被引规律,不仅是论文未被引现象研究的重要内容,也有利于扩展基于内容的引文分析方法范畴。[方法/过程]以 CSSCI 作为来源数据库,以图书馆情报与文献学为样本学科,依据该学科学者的 h 指数分布特征随机选择 200 名学者作为样本对象,下载其 1998 – 2015 年的所有被收录论文数据;下载样本学科 1998 – 2015 年的所有收录论文数据,并离析出对应被引论文、高被引论文的相关数据;以 6 年为时间窗口,将发表后 1 – 3 年内被引的论文定义为被引论文,其余的为未被引论文;析取未被引论文、被引论文、学科整体论文及高被引论文的关键词,按关键词频数从高到低排序,选取排序前 50 的关键词构成关键词向量,计算关键词向量的内积、欧几里得长度和余弦相似度。[结果/结论]图书馆情报与文献学领域在 21 世纪初形成较为稳定的研究内容体系,其未被引论文与学科整体论文、被引论文、高被引论文的内容相似度都较低,表明研究内容对论文未被引有重要影响。

关键词: 未被引论文 零被引论文 图书馆情报与文献学 论文内容特征 论文关键词 向量空间模型
分类号: G250.252 G301
DOI:10.13266/j.issn.0252 – 3116.2018.04.002

1 引言

科学交流系统是以相关为基础的知识生产、交流与利用过程,而引用关系是关联这些环节的核心要素,引文分布分析则成为科学计量学的核心研究主题[1]。然而,目前引文分布研究更多聚焦于其核心区域,较少关注未被引或引用较少的长尾部分,但任何学科或任何地方出版的论文中,普遍存在一些从未受到任何引用的论文[2]。因此,引文分布规律的完整认知不可或缺地包含未被引现象研究。

关于论文未被引问题很早就受到学者关注,但大规模研究则是在 20 世纪 90 年代。目前关于未被引现象研究主要集中在以下 3 个方面[3-4]:论文未被引现象的测度指标设计及模型研究,使用的测度指标主要是未被引率;论文未被引现象的影响因素,主要聚焦于参考文献数量、作者数量、论文长度、论文所属学科、国际合作关系、学科跨度、发文机构以及刊载期刊等方面对论文是否被引的影响关联程度;论文未被引的分布特征与分布规律,包括特定期刊分布、特定学科领域分布、国家分布等,尤其集中于特定期刊分布以及这些分布特征与规律在实践中的应用探索。

然而,无论是引文分析还是未被引现象研究,通常只是从形式特征来展开分析。随着基于内容引文分析方法的出现[5],如何在未被引现象研究中引入基于内容的方法显得尤其重要。本研究尝试从未被引文献与其它文献(如学科整体文献、引用文献及高被引文献等)的内容差异来探究文献内容对未被引的影响,为未被引现象研究引入内容分析方法提供一种可能,同时扩展基于内容的引文分析方法范畴,丰富情报学的研究内容。

2 相关研究工作

论文未被引作为科研成果的一个重要特征在 20 世纪 50 年代就受到关注,E. Garfield 指出论文未被引的影响因素很多[6]。大部分未被引影响因素研究都集中在文献外部特征[7-8],如作者数量、关键词数量、标题词数量、参考文献数量、期刊年龄、期刊价格等,尤其是从特定学科期刊入手的研究相对较多[9-10]。即使关

作者简介:韩毅(ORCID:0000 – 0001 – 7021 – 3229),教授,博士,博士生导师,E-mail:hanyi72@swu.edu.cn;伍玉,硕士研究生;申东阳,硕士研究生;况书梅,硕士研究生;袁庆,硕士研究生。
收稿日期:2017 – 08 – 31 修回日期:2017 – 12 – 02 本文起止页码:14 – 20 本文责任编辑:王传清

注关键词、标题词这些反映文献内容特征的因素，也仅是从数量特征切入，较少从内容视角来探讨对未被引的影响程度。

但是也有少量学者关注到内容因素对未被引的影响。温芳芳通过零被引论文与高被引论文的比较，基于关键词共现网络发现高被引论文更关注研究热点，而未被引论文的研究热点关键词绝对频次低且各关键词的分布频率较为平均[11]；钟镇以高频叙词作为研究热点，比较9种农业经济与政策SCI源刊高被引论文与零被引论文的选题，研究结果显示绝大多数期刊的高被引论文在总体上都有着比零被引论文更高的热点叙词分配率[12]；高继平、潘云涛和武夷山以光谱学领域未被引论文作为对象，通过主题分析发现未被引论文的主题分布与学科整体的热点主题分布有所不同[13]；李江在评述科学"睡美人"与"昙花一现"文献时提到，文献主题对文献的引用趋势存在一定程度的影响，从另一个侧面反映出未被引受到文献内容的影响[14]。

已有这些与内容相关的未被引研究，要么是比较未被引与高被引的关键词频率分布，要么分析未被引与高被引的热点叙词分配率，还有的仅从定性视角来说明两者在内容上的差异性，没有从计算角度来分析未被引与高被引的相异性。本研究期望通过特定样本学科的未被引论文与高被引论文、学科整体论文、学科被引论文的内容差异性计算来探讨内容特征对论文未被引的影响。

3 数据与方法

3.1 数据获取

本研究以图书馆情报与文献学为样本学科，于2016年11月在CSSCI中采集样本学科的相关数据，由于2016年引文数据不全，因此所有数据仅限于1998－2015年发表的期刊论文。

具体的数据收集策略如下：①综合图书馆情报与文献学学者们的h指数特征[15]，随机抽取200名学者作为样本对象；②从CSSCI获取以样本学者为第一作者、于1998－2015年期间发表期刊论文的详细信息，以txt文件方式存储下载的各数据项；③从CSSCI获取收录的图书馆情报与文献学学科1998－2015年所有论文数据，获取所有论文的关键词与发文后被引数据，以txt文件方式存储下载的各数据项；④从CSSCI获取收录的图书馆情报与文献学学科1998－2015年所有论文数据的引文数据，遴选出各年的高被引论文（根据选择时点的统计数据，利用 $n=0.749\sqrt{n_{max}}$ 计算出高被引论文的最低引用数阈值，进而确定高被引论文集合），以其为对象获取这些论文的关键词与被引数据，以txt文件方式存储下载的各数据项。

3.2 研究方法

关于文献内容的识别与统计，最好采用标题词表或叙词表进行规范。然而，中文环境下缺乏特定学科的标题词表或叙词表支持，而论文关键词与其内容紧密相连，是论文内容的最直观体现，也是论文中显著的标注数据，因此本研究选择关键词作为论文内容的测量项。

分析发现，图书馆情报与文献学领域论文被引高峰期在发表后2－3年。考虑到引用延迟现象，本研究以6年为统计时间窗口，以3年为被引分界窗口，将发表后1－3年内被引的论文定义为被引论文，其余的论文定义为未被引论文。

提取样本对象的被引论文、未被引论文、学科整体论文、高被引论文集合的关键词，按频次从高到低进行排序，选择前50个关键词构成词向量空间，以此为基础分别计算出各个年度被引论文、未被引论文、学科整体论文、高被引论文之间的相似度，探究他们之间的内容相关性，以此探索与揭示内容因素对未被引的影响程度。

3.2.1 关键词权重计算　通过CSSCI获取1998－2015年图书馆情报与文献学领域各年的发文，以3年为一个周期，利用Citespace提取每个周期的学科关键词并计算其词频。假设学科整体论文集有t个关键词，其对应的权重分别是 g_1、g_2、…、g_t，关键词权重 g_k 计算公式如下：

$$g_k=\frac{d_k}{N_g}\ (k=1,2,\cdots,t)$$

其中：d_k 为表示关键词k的词频，N_g 为表示论文总数量，其值大小在[0,1]区间中。

同理，以获取的图书馆情报与文献学领域200名学者1998－2015年发文为样本，分年进行数据处理。使用自编软件分别提取各年被引与未被引关键词及其词频，其中被引论文关键词及其词频从论文发表后1至3年内有被引的论文中提取，而未被引论文关键词及其词频的提取则在其余论文中提取。采用类似的方法分别计算被引论文关键词权重 c_i 与未被引论文关键词权重 nc_j。其计算公式如下：

$$c_i=\frac{d_i}{N_c}\ (i=1,2,\cdots,m),\ nc_j=\frac{d_j}{N_{nc}}\ (j=1,2,\cdots,n)$$

其中：d_i 为关键词i的词频，N_c 为计量窗口被引文

章总数量，d_j 为关键词 j 的词频，N_{nc} 为计量窗口未被引文章总数量，其值大小在[0,1]区间中。

3.2.2 关键词向量内积计算 被引论文、未被引论文、高被引引论文、学科整体论文对应的关键词向量分别用 $\underset{V}{\rightarrow}(c)$、$\underset{V}{\rightarrow}(nc)$、$\underset{V}{\rightarrow}(hc)$、$\underset{V}{\rightarrow}(g)$ 表示，则可计算向量两两之间的内积来表示关键词向量相关度，即在内容上的相似程度。

计算关键词向量内积本质上是将两个向量空间中的词项进行合并。计算时，各词项可能出现两种情况：一是在被引（未被引）论文或学科整体论文的关键词向量权重为0；二是两个向量中均不为0。前一种情况下其计算结果取值为0，后一种情况下的内积计算略显复杂。下面以被引关键词向量与学科关键词向量为例给出定义式。

设被引论文关键词集与学科整体论文关键词集存在 e 个共同语词，g_1、g_2、…、g_e 与 c_1、c_2、…、c_e 为共同语词分别在学科整体、被引论文中的权重，则被引论文与学科整体论文的关键词向量内积计算公式为：

$$\underset{V}{\rightarrow}(c) * \underset{V}{\rightarrow}(g) = \sum_{i=1}^{e} c_i * g_i$$

同理，可定义其他类别关键词向量的内积。

3.2.3 关键词向量的余弦相似度计算 各年学科整体论文、被引论文、未被引论文、高被引论文所涉及的关键词数量较多，导致各关键词词频相对较小、权重数值较小，最终计算所得的内积值较小。为此，本研究进一步计算关键词向量的余弦相似度，通过余弦相似度的分母做标准化处理，以消除不同样本集合容量大小的影响。

根据余弦相似度计算原理，可定义被引论文与学科整体论文的关键词向量余弦相似度计算公式为：

$$\mathrm{sim}(c,g) = \frac{\underset{V}{\rightarrow}(c) * \underset{V}{\rightarrow}(g)}{|\underset{V}{\rightarrow}(c)|\,|\underset{V}{\rightarrow}(g)|}$$

同理可定义其他任意两类数据集的余弦相似度计算公式。从计算公式可见，关键词向量的余弦相似度计算公式的分子是向量内积，反映了共同关键词数量的多少，其值越大，内容的相似度超高；分母是关键词向量欧几里得长度的乘积，而欧几里得长度反映了对应向量的关键词数量多少，其值越大表明学科的研究内容较为集中，否则其研究内容较为分散。因此，还可利用欧几里得长度来探索样本对象在研究内容上的集中程度。

4 研究结果

4.1 样本学科整体的内容特征

样本学科的整体内容是论文被引、未被引、高被引现象的基本环境，也是引用行为的基本约束条件，因而有必要揭示样本学科的整体内容特征。

为了刻画样本学科整体在研究内容上的延续性与差异性，提取样本学科1998－2015年所有发文的关键词，计算各年发文内容与其后各年发文内容的关键词向量内积并可视化展示（见图1）。为了保证每个时间序列都有足够的数据呈现，只计算到2011年序列。如2011年序列，则计算2011年与其后的2012－2015年的向量内积，2012年为2011年序列的1年后，2013年为2年后，以此类推。从图1中可见：随着时间的演化，各年与其后各年的内容向量相似度逐渐减小，相邻年度间的关键词向量相似度较大，既表明图书馆情报与文献学科在研究内容上的时间延续性，也表明研究内容的不断创新与拓展。

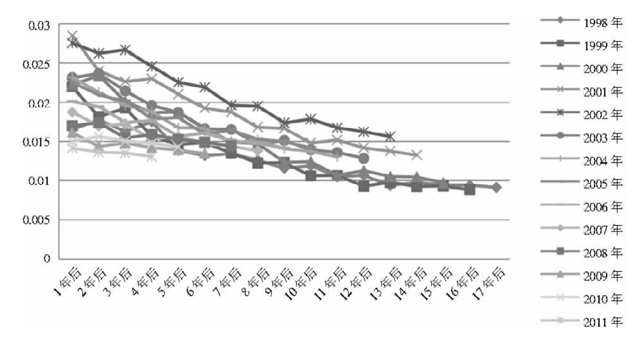


图1 1998－2011年各年发文与其后各年发文的关键词向量内积

以各年与其后各年关键词向量内积为对象绘制成时间维度的演化趋势图可见，基本趋势是相似度先上升后逐年下降，其中2002年学科研究内容与其他年份相似度最高（见图2）。据此可以推断，2002年左右图书馆情报与文献学科的研究内容既是前面年份的一种凝聚，从较为分散向较为集中方向发展；其后图书馆情报与文献学科的研究内容向外拓展发散，表现出不断分化的发展态势。

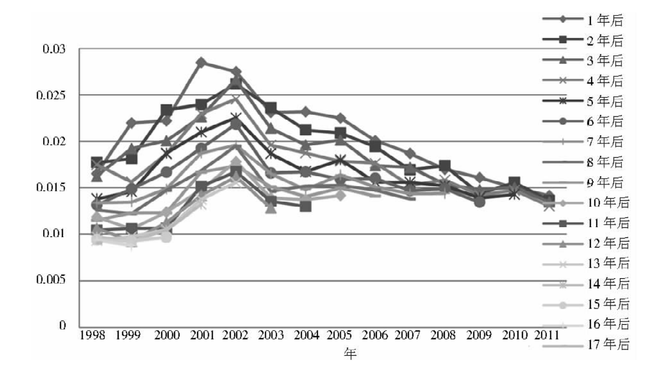


图2 1998－2011年的各年内容向量内积演化趋势

图3是样本学科各年研究论文内容的欧几里得长度计算结果，可在一定程度上反映样本学科各年的研究内容凝聚度，其值越大则内容凝聚度越大，反之亦然。由图3可见：2002年前后，图书馆情报与文献学科的主要研究内容较为集中；之前的各年份在研究内容上有逐渐凝聚的趋势，之后的各年份在研究内容上有逐渐拓展分化的态势。因此，2002年左右可能是图书馆情报与文献学科研究内容的转折点。

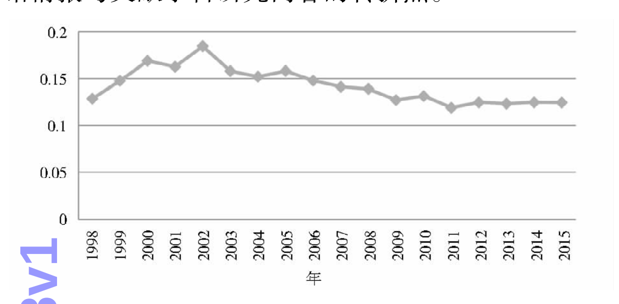


图3　1998－2015年各年学科内容的欧几里得长度

图4是依据1998－2015年各年间内容向量的余弦相似度绘制成的演化趋势图。由图4可知：所有序列大致可分为两个部分：1998－2000年序列为一组，其它序列为一组，前者的余弦相似度值比后者低，表明前者在研究内容上较为分散，后者的研究内容更为凝聚；1998－2000年序列组的特征是其后1－3年的研究内容向量余弦相似度较高，3年以后的向量余弦相似度较低，表明其研究内容有不断扩展的态势；2001－2011年序列组其后各年的研究内容向量余弦相似度没有明显的变化趋势，表明2001年以后学科的研究内容保持相对稳定，但也体现了一定的拓展分化趋势。

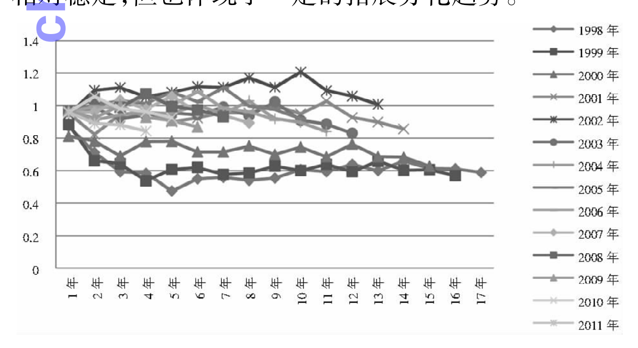


图4　1998－2011年各年间研究内容的向量余弦相似度

1998－2011年与其后各年不等年限的内容向量余弦相似度演化发展趋势见图5。从演化趋势可以看到：1998年与其后1－2年的内容向量余弦相似度较高，1999年与其后1年的向量余弦相似度较高，1998年、1999年、2000年与2001年及以后各年的内容向量余弦相似度较低，而2002－2015年的内容向量余弦相似度都保持在较高且稳定的水平，特别是2002年与其后各年的内容向量余弦相似度都保持最高水平。这说明图书馆情报与文献学科1998－2000年的研究内容较为稳定，而2001年左右研究内容出现了一定变化，2002年以后各年学科研究内容又保持在较为稳定的状态，说明2001年左右可能是图书馆情报与文献学科研究内容发生改变的转折点。

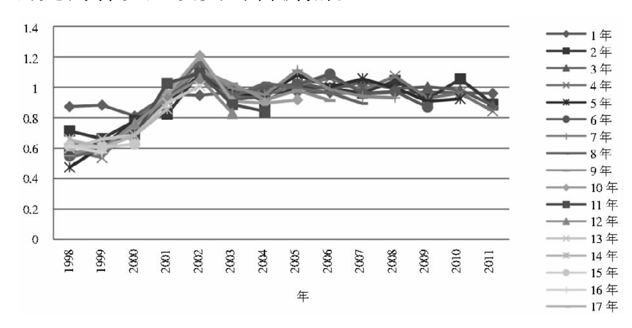


图5　1998－2011年各年间研究内容向量余弦相似度的演化特征

综合上述研究结果可以看到：图书馆情报与文献学科在整个研究内容上，21世纪初叶形成了较为稳定的内容体系，之前较为分散的研究内容逐渐凝聚成共识，其后在共识基础上有一定程度的拓展，2001、2002年可能是整个学科发展的关键节点。分析其中的原因，大致有以下两个方面：①图书馆情报与文献学科经过几十年的发展逐渐走向成熟，科学共同体对于学科的基本内容有了较为一致的看法；②20世纪末的大规模互联网应用带来了信息环境的深刻变化，为图书馆情报与文献学科的发展开拓了新的领地，研究范围不断扩展，以关键词为表征的内容也逐渐走向发散。

4.2　被引论文和未被引论文的内容差异特征

各年度被引论文、未被引论文与学科整体论文的关键词向量内积变化趋势见图6。从图6可以看到：变化趋势基本相似，被引论文与学科整体论文的向量相似度总体上高于未被引论文与学科整体论文的向量相似度；早期的相似度差距较大，随着时间的发展差异越来越小，甚至在2011和2012年未被引论文与学科整体论文的向量相似度超过被引论文与学科整体论文的相似度，表明在学科整体背景下，研究内容的同质替代性更强，竞争更加激烈；被引论文、未被引论文与学科整体论文的内容相似度整体上有下降的趋势，表明研究内容的发展越来越分散，研究范围在不断扩展。分析学者样本数据与学科整体数据发现，近年学科整体发文量逐渐增大，内容涉及的范围越来越广，表现为关键词数量增多，而本研究所选取学者样本数据各年数

量相对平衡，这可能是样本学者论文与学科整体论文的相似度逐年下降的原因之一。



图 6 1998－2013 年被引、未被引论文与学科整体论文内容向量的内积演化趋势

图 7 是学科整体论文、被引论文、未被引论文的向量欧几里得长度情况。从图 7 可以发现：被引论文与未被引论文相对于学科整体论文的内容集中度相对较高，即学科整体论文在内容上有扩展趋势，而被引、未被引论文内容相对凝聚。这一方面与图书馆情报与文献学科不断拓展研究边界有关，另一方面也与样本数据有关。由于学科整体论文的欧几里得长度计算来自收集的学科核心期刊整体载文，论文数量巨大，计算所得欧几里得长度相对较小；而所选取样本数据来自学界部分学者，相较而言论文数量较小，导致关键词权重较大，计算所得欧几里得长度较相对较大。

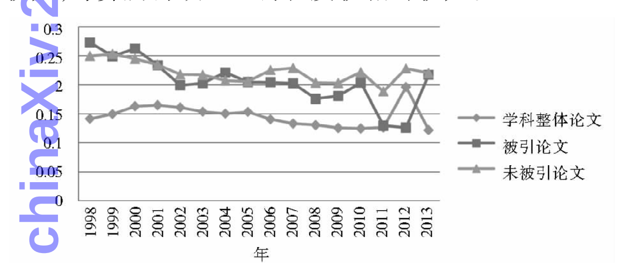


图 7 1998－2013 年各年学科整体论文、被引论文、未被引论文内容向量的欧几里得长度

图 8 是 1998－2013 年被引论文、未被引论文关键词向量的余弦相似度情况。之所以把这两者单独呈现，原因在于被引论文与未被引论文是以选择的 200 名样本学者所发表的论文为基础计算的，其绝对值相对于学科整体论文的值较小，计算值相对较低。由图 8 可见，整体上两者的向量相似度逐年下降，即被引论文与未被引论文两者间内容差距有增大的趋势。1998－2000 年期间在内容相似性上呈逐步上升态势，这与前面学科整体内容在早期的逐步凝聚有关系；其后基本上呈逐步下降态势，表明被引论文与未被引论文在研究内容上有渐行渐远的趋势。这些说明，研究内容对论文未被引有显著影响。

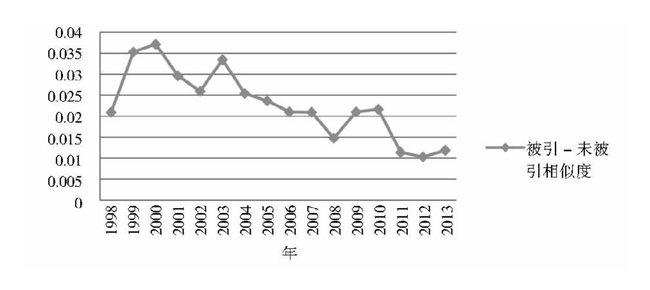


图 8 1998－2013 年各年被引论文与未被引论文内容向量的余弦相似度

图 9 是 1998－2013 年各年被引论文－学科整体论文、未被引论文－学科整体论文、高被引论文－未被引论文的内容关键词向量的余弦相似度情况。从图 9 中可以发现：3 条曲线的基本走势大致相似，表明各年度的核心关键词具有较高的相似性；各年被引－学科论文的内容向量余弦相似度比未被引－学科论文的内容向量余弦相似度要高，表明被引论文的内容与学科整体论文的共同核心关键词更多，与学科在整体上保持了较为一致的发展态势，而未被引论文与学科整体论文在共同核心关键词相对较少，与学科发展的整体热点有所背离；高被引－未被引论文的内容向量余弦相似度最低，表明两者间的共同核心关键词数量较少，在内容上的差异较大。

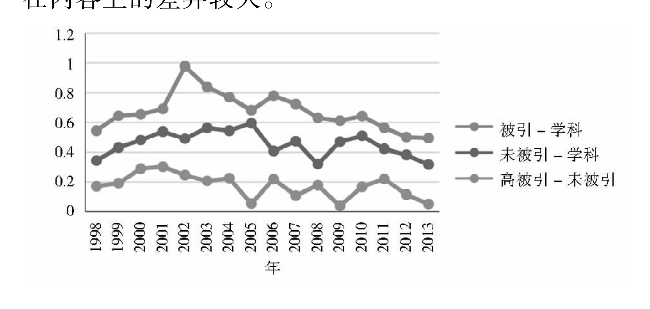


图 9 1998－2013 年被引－学科、未被引－学科、高被引－未被引论文的内容向量余弦相似度

通常来讲，学科整体论文代表着学科发展全局，其高频关键词代表着学科的整体研究内容取向，是学科延续性发展的核心驱动力；被引论文与学科整体论文在内容关联性上相较于未被引论文与学科整体论文的相关度更高，被引论文与未被引论文在内容关联性上较低，一方面说明未被引论文研究内容与被引论文、学科整体论文有差别，另一方面也表明论文内容与当前学科基本发展态势越贴近则越容易较快获得引用；高被引论文代表着学科领域的知识内核，是学科知识体系的核心力量，在领域知识传递中起着关键性作用，其核心关键词代表着领域的基本发展方向，未被引论文与高被引论文关键词内容的较低相似度表明未被引论

文所包含的知识游离于学科知识内核之外，对学科知识内核发展贡献相对较低，因而其获得引用的可能性相对降低。

5 结论与讨论

5.1 主要发现

5.1.1 21世纪初叶是图书馆情报与文献学领域研究内容的转折点　样本领域计量时间范畴内的研究内容特征是探索未被引论文内容影响因素的基本约束条件。无论是向量内积值、欧几里得值还是余弦相似度值都表明图书馆情报与文献学科在21世纪初形成较为稳定的研究内容体系，从之前的发散状态向凝聚状态的演化，再到从之后的凝聚状态向发散状态的演化。这表明学科领域的研究内容可遵循"发散－凝聚－发散－凝聚"的基本模式，通过发散来拓展学科的研究内容、研究范畴甚至研究边界，通过凝聚来凝练学科的知识内核与核心知识共识，基于这些嬗变过程不断地夯实学科的知识硬核，体现学科发展的稳定性、延续性与拓展性。

5.1.2 内容是图书馆情报与文献学领域未被引现象的重要影响因素　无论是各年度被引论文、未被引论文与学科整体论文的关键词向量内积值还是学科整体论文、被引论文、未被引论文间的向量欧几里得长度值，无论是被引论文、未被引论文关键词向量的余弦相似度值还是各年被引论文－学科整体论文、未被引论文－学科整体论文、高被引论文－未被引论文关键词向量的余弦相似度值都表明未被引论文与代表背景内容的学科整体论文在核心关键内容上有所偏离，未被引论文与代表受到关注的被引论文在关注内容上存在一定差异。即当论文内容与学科整体核心内容、学科受关注的核心内容吻合度较高时，其获得引用的几率就会增加；而当文献内容与这些核心内容偏离较大时，其不被引用的几率可能增加。因此，文献未被引现象受到研究内容的强烈影响，内容是论文未被引的重要影响因素。

5.2 对比分析

研究内容对论文未被引的影响在诸多研究都有提及[6,13-14]，但具体是怎样的影响形式及影响特征并没有得到详细的论证；一些研究从高被引论文与未被引论文间的关键词分布来讨论研究内容对论文未被引的影响，仅是通过对比两类不同数据集中的关键词分布差异[11-12]，没有对这些差异的具体程度进行计算。

本研究以关键词作为论文内容的直观反映，选择各类别的高频关键词构建内容向量空间，从相似性计算来探讨论文内容对未被引的影响程度，能够更准确地量化不同类别间的差异程度。在实践中，首先，探讨了样本学科在计量窗口中研究内容的整体特征，它构成了论文被引与论文未被引研究的知识背景画面；其次，通过计算被引论文、未被引论文与学科整体论文的核心关键词向量内积、欧几里得长度和余弦相似度，尤其是高被引论文与未被引论文的余弦相似度，从数量特征上证实了研究内容对论文未被引有重要影响。

6 结语

本研究以论文研究内容为突破口，以关键词作为内容的测度项目，通过引入向量空间模型分析未被引论文、被引论文、学科整体论文及高被引论文在研究研究内容上的差异，证实了未被引论文在研究内容关注点与学科整体论文、被引论文、高被引论文都有明显差异，表明研究内容对论文未被引有显著影响。

本研究结论由CSSCI所收录的图书馆情报与文献学样本学科的数据获得，其结论的普适性与推广性需要进一步增加中文环境下其它样本学科的相关数据、非中文环境的样本学科数据来进一步验证；同时，关键词作为作者给出的描述文献内容的非规范化语词，对于相同内容不同作者可能会因使用习惯不同、知识结构差异而给出不同的描述结果，而且即使是相同关键词其具体内容的描述侧重点也可能有较大差异，尽管通过大量的数据可能稀释这种影响，但这种影响显然是无法完全避免的，因此需要寻求规范化标题词表或叙词表来进一步证明本文方法的有效性；以学者h指数特征来随机选择样本对象且只选择了样本对象以第一作者身份出现的论文数据，尽管采用无关标志可以在一定程度上避免对内容选择的干扰，但实践上可能会引发系统性误差，这种情况需要通过分层抽样来进一步提高误差控制精度；向量长度选择具有较大的经验色彩，是否存在最优化的向量长度取值也是值得进一步探索的问题；样本数据中，被引论文与未被引论文的时间窗阈值选择是根据引用峰值年确定的，如果采用其它方法（如半衰期）来划分时间窗是否会得到不同的结果，也是值得进一步研究的问题。所有这些问题与局限都需要在未来研究中需要进一步探索，以期得到更为一般性的结论。

参考文献：

[1] STRINGER M J, PARDO M S, AMARAL L A N. Statistical validation of a global model for the distribution of the ultimate number

of citation accrued by papers published in a scientific journal [J]. Journal of the American Society for Information Science, 2010, 61(7): 1377-1385.

[2] BURRELL Q L. Will this paper ever be cited? [J]. Journal of the American Society for Information Science and Technology, 2002, 53(3): 232-235.

[3] 石磊. 期刊论文零被引现象实证研究[D]. 蚌埠: 安徽财经大学, 2016.

[4] 胡泽文,武夷山. 零被引研究文献综述[J]. 情报学报, 2015, 34(2): 213-224.

[5] ZHANG G, DING Y, MILOJEVIC S. Citation content analysis (CCA): a framework for syntactic and semantic analysis of citation content[J]. Journal of the American Society for Information Science and Technology, 2013, 64(7): 1490-1503.

[6] GARFIELD E. To be an uncited scientist is no cause for shame [J]. The scientist, 1991, 5(6): 12.

[7] STERN R E. Uncitedness in the biomedical literature [J]. Journal of the American Society for Information Science, 1990, 41(3): 193-196.

[8] ROUSSEAU R. Why am I not cited, or why are multi-authored papers more cited than others [J]. Journal of documentation, 1992, 48(1): 79-80.

[9] 刘武英, 张薇, 刘影梅. 学术期刊中的零被引论文特征分析——以编辑出版类核心期刊为例[J]. 中国科技期刊研究, 2015, 26(9): 987-991.

[10] 郭亿华. 地理学中文核心期刊零被引论文特征分析[J]. 中国科技期刊研究, 2016, 27(10): 1094-1099.

[11] 温芳芳. 我国情报学期刊论文零被引的成因及影响因素探析[J]. 情报理论与实践, 2016, 39(4): 27-31, 26.

[12] 钟镇. 基于叙词分析的高被引与零被引论文选题差异研究——兼谈标准叙词索引与引文索引数据的混合应用[J]. 情报学报, 2014, 34(11): 1185-1193.

[13] 高继平, 潘云涛, 武夷山. 零被引论文的形成因素分析——以光谱学领域零被引论文的国家、机构和主题分布为例[J]. 科技导报, 2015, 33(8): 112-119.

[14] 李江. 科学中的"睡美人"与"昙花一现"现象评述[J]. 大学图书馆学报, 2016, 34(3): 38-43.

[15] 韩毅, 夏慧. 时间因素视角下科研人员评价的 Pt 指数研究[J]. 中国图书馆学报, 2015, 41(6): 73-85.

作者贡献说明:

韩毅: 整体研究设计与规划, 论文修改;
伍玉: 数据搜集、整理与分析, 论文撰写;
申东阳: 程序编写, 数据搜集;
况书梅: 参与数据搜集、整理与分析;
袁庆: 参与数据搜集、整理与分析。

Part Ⅱ of the Exploration on Uncited Papers in Chinese: The Influences of Content Features Based on Keywords in Paper - A Case Study of Library and Information Science

Han Yi Wu Yu Shen Dongyang Kuang Shumei Yuan Qing

College of Computer and Information Science, Southwest University, Chongqing 400715

Abstract: [**Purpose/significance**] It is of great importance to study the law of uncitednessfrom content differences, which is not only the important content in studying uncitedness phenomena, but helps to expand the boundary of citation content analysis. [**Method/process**] CSSCI was selected as the source database, and library and information science was chosen as the sample source. According to the features of the h index, 200 scholars were selected randomly as samples, and their related data, recorded in CSSCI from 1998 to 2015, were downloaded. All the collected data of library and information science from 1998 to 2015 were downloaded, and their relevant data about cited papers and highly cited papers were extracted. Taking 6 years as time window, the papers cited in 1 to 3 years were defined as cited papers, and the others as uncited papers. The key words of uncited papers, cited papers, all discipline papers and highly cited papers were taken and listed according to keyword frequencies from high to low, first 50 keywords were selected to be keywords vector, and their inner product, Euclidean length and cosine similarity were calculated respectively. [**Result/conclusion**] The results have showed that: the research content of library and information science has been probably stable at the beginning of 21st century; the content similarity between uncited papers and all discipline papers, cited papers, and highly cited papers are lower, which means the research content has a significant effect on uncited papers.

Keywords: uncited articles non-cited articles library and information science content features in article keywords in article vector space model